DOCUMENT RESUME

ED 200 906 CS 005 996

AUTHOR Hayford, Paul D.; Salter, Ruth
TITLE Rule-Based Measures of Literal Comprehension.
PUB DATE Mar 78
NOTE 52p.; Paper presented at the Annual Meeting of the American Educational Research Association (Toronto, Canada, March 27-31, 1978).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Measurement Techniques; *Reading Comprehension; *Reading Tests; *Test Construction; *Test Reliability; *Test Theory; *Test Validity

ABSTRACT

Reading comprehension involves a number of distinctly different intellectual skills that can be assessed if the proper techniques are employed. As part of a reading assessment system, two measures of literal comprehension were developed: the Literal Comprehension Details Test (LCDT) and the Paraphrase Reading Test (PRT). Both the LCDT and the PRT assume the possession of visual and phonetic skills prerequisite to comprehension but do not presume to assess those skills necessary for processing beyond the apprehension of the explicit meaning of the text. The LCDT was conceived as a battery of test passages, scaled by difficulty level, with accompanying rule-based items for measuring literal comprehension, whereas the PRT was developed for use as a criterion measure of literal comprehension. The passages for the PRT are the same passages that are used in the Multiple-Choice Cloze Exercises. The basic difference between the PRT and LCDT items, then, is that the PRT items involve paraphrase. A study of validity indicates that the LCDT has high face validity as a measure of literal comprehension and that the face validity for the PRT is higher than that for the PCDT. (Appendixes include rules for constructing wh-detail items and rules for constructing items for paraphrase reading tests.) (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200906

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

RULE-BASED MEASURES OF LITERAL COMPREHENSION*

Paul D. Hayford and Ruth Salter

March 1978

Bureau of School and Cultural Research
New York State Education Department
Albany, New York 12234

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Paul D. Hayford
Ruth Salter
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PRINTED IN USA

CS005996

*Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

The assessment of reading comprehension is a significant issue for educators and educational researchers wishing to respond to the widespread concern over literacy and illiteracy. This paper will make no pretensions to solving the problem, but will try to contribute practical information to the continuing discussion. The paper will describe the rationale, construction, and analysis of two tests developed to measure literal comprehension, the Literal Comprehension Details Test (LCDT) and the Paraphrase Reading Test (PRT). Both tests were developed by the Bureau of School and Cultural Research of the New York State Education Department.

A Measurement Problem

One of the fundamental problems in the measurement of any intellectual (non-observable) process, skill, or ability is the degree of relationship between the intellectual phenomenon and the instrument or method used to measure it. In the measurement of mathematical skills, for example, the connection is often quite apparent. The relationship between the ability to add and the correct answering of addition examples is one of identity; there can be no better evidence of the possession of addition ability than the correct answering of addition examples. Similarly, the relationship between the ability to recall dates of historical events and the provision of such dates for a given list of events is also identity. But in the measurement of the processes, skills, or abilities involved in reading comprehension, the relationships between the intellectual phenomena and the techniques used to assess them are seldom identity.

For a given reading selection or portion of printed discourse, reading comprehension may involve a wide variety of skills. Comprehension of a particular prose passage, for example, may involve skills as disparate as those related to grammar and vocabulary, inference, propositional logic,

critical reasoning, metaphoric interpretation, and rhetorical analysis. Assessment of a reader's comprehension of such a passage could be attempted in a number of ways, but no single method could tap all of the skills involved in comprehending the passage.

A few brief illustrations may suggest the breadth of skills and the consequent types of assessment potentially involved even in a relatively short and uncomplex passage:

> Big Jim had to duck his head to get through the entrance. Everybody else had to do the same.

If we wanted to assess a reader's comprehension of this passage, our assessment procedure would be governed by our assessment purposes. If the reader whose comprehension we wanted to assess was a pupil in the primary grades, we might use items like the following:

> Who had to duck his head to get through the entrance? (Verbatim wh- detail item stem)
>
> Big Jim had to duck his head to get through the _____. (verbatim completion item)
>
> What did Big Jim need to do to pass through the opening? (paraphrase wh- detail item stem)

But if we were interested in assessing a more experienced reader's inferential ability, we might use the following method:

> Does the above passage contain sufficient information to enable us to tell why Big Jim and everybody else had to stoop to get through the entrance? Answer in complete sentences.

As these examples have tried to suggest, the assessment of different types of skills or abilities requires the use of different types of assessment devices. Measurement of some kinds of comprehension, at the

explicit level, for example, may be accomplished by objective items like the three given above. But it would be difficult, if not impossible, to assess certain higher-level comprehension skills by means of objective items. For instance, an objective item could be constructed to assess inferential skills similar to those measured by the sample item above which required a response in complete sentences:

> We cannot tell why Big Jim and everybody else had to duck their heads to get through the entrance because . . .

A possible correct response might suggest the lack of sufficient context (Big Jim and the others might be giants, for example, or they might be normal-sized people in a land like Lilliput). But such an item would not be measuring the same skills as the item which required a complete-sentence answer. To select a valid inference from a list of inferences of varying degrees of plausibility is not equivalent to drawing a valid inference.

The insistence on this distinction may seem like nitpicking, but the assumption that such different items measure the same things yields unhappy consequences. To wit, either there are no distinctions to be made among the intellectual (as opposed to the physical) skills related to reading comprehension, or there is no viable method of distinguishing among such varied skills as are related to reading comprehension. If either of these assumptions is made (and one or the other must be if we accept the initial premise in this paragraph), then it follows that any reading test which presupposes decoding will be as good or as useful as any other reading test. Essentially, this conclusion would render the use of reading tests futile, for results of such tests would be largely uninterpretable.

Pursuing this argument further, if we are unwilling to acquiesce in the futility implicit in the previous line of reasoning, then we must contend

that reading comprehension involves a number of distinguishably different intellectual skills and that these different skills can be assessed if the proper techniques are employed. The aim of this paper is to describe and evaluate an attempt at identifying skills or ranges of skills involved in reading comprehension and the accompanying techniques developed to assess the skills so identified.

As mentioned previously, the two tests under discussion, the Literal Comprehension Details Test (LCDT) and the Paraphrase Reading Test (PRT), were both developed as measures of literal comprehension. The LCDT, developed during 1974 and 1975, was produced as part of a reading assessment system conceived as comprehensive and multi-faceted. The system would have included a wide variety of test items for the measurement of a broad spectrum of reading skills. Though the system was never completed in the form in which it was conceptualized, the LCDT was produced as one of the literal comprehension components. The PRT was constructed during 1976 and 1977 expressly as a criterion measure of literal comprehension for use in the construct validation of the Multiple-Choice Cloze (MCC) Exercises, another measure developed by the Bureau of School and Cultural Research to assess literal comprehension.

Defining Literal Comprehension

Detailed analysis of literal comprehension and extended discussion attempting to define the term and establish a construct of literal comprehension are recorded elsewhere ("Construct Validation of Multiple-Choice Cloze Exercises," 1977; Kidder, 1976; Schuder, Kidder, & O'Reilly, 1976; O'Reilly, Schuder, & Kidder, 1976). Rather than attempting or repeating prior attempts at a precise technical definition of literal comprehension, the present discussion will try to describe in a nontechnical way the limits of the range of skills

involved in literal comprehension.

The term literal comprehension can entail two separate but related concepts. Literal comprehension can refer to the skill or process the application of which or during which a reader apprehends the literal meaning of discourse, the meaning of discourse at its literal level. Or literal comprehension can signify the result or product of the process of apprehending the literal meaning of discourse.

It is clear that the acquisition of the product literal comprehension implies that the process of apprehending meaning at the literal level of discourse has occurred. However, as suggested previously, the process of literal comprehension is unmeasurable because it is an unobservable, intellectual process. But since there cannot be a literal comprehension "product" independent of a literal comprehension "process," the measurement of the product literal comprehension is a direct indicator of the efficiency of the process literal comprehension.

All discourse which has clear and unambiguous meaning possesses a literal level. Without a literal level of meaning, discourse could have no determinate complex, inferential, or higher level of meaning. The literal level of meaning, then, is the foundation upon which all other meanings rest.

The reply may be made that many words, and even many sentences, can have multiple meanings, and that these multiple meanings refute the contention that a literal level underlies all other meaning. But such a position neglects the context which discourse provides and which limits and excludes many possible meanings in favor of a single meaning. (No words or sentences with communicative purpose occur independent of context.) The contexts of discourse disambiguate potentially ambiguous words and sentences. Consider the following sentence:

We had a ball.

As it stands, independent of any surrounding context, the sentence is ambiguous. Its potential meanings include, for example, we held a formal dance, we had in our possession a round object used in games, or we experienced a time of enjoyment. But given context, the ambiguity is dissipated:

> We all wanted to play baseball. We had a ball. But we had no bat. We were frustrated.

In this example, the context clearly excludes the first and third of the potential meanings noted above (as well as any others) and specifies the second meaning. Indeed, from the context we know not only that ball refers to a round object used in games, but also that the ball is the kind used in baseball games: a baseball.

Context, then, which discourse (as opposed to individual words or sentences) provides, does exclude and disambiguate, so that where there is clearly specifiable meaning there is a literal level of meaning.

To anticipate one further objection, it may be argued that some writers, and especially modern writers, have attempted, in poetry and in prose, to suggest aspects of their experience which provoke anxiety or frustration or seem incoherent. It may be asserted that, in conveying such impressions, writers produce passages which have no literal level of meaning. But the obvious reply is that there is a distinction to be made between the appearance or suggestion of incoherence (consciously rendered and controlled) and incoherence itself. When a writer's efforts result in the latter effect--incoherence--we no longer accord him the title of writer; we merely observe that he has (not permanently, we hope) lapsed into incoherence. He has, in short, failed clearly to communicate; if he has aimed at producing multiple levels of meaning or interpretation, he has, through the absence

of a foundation for those meanings or interpretations, missed his mark. He has not provided a literal level of meaning.

Practically speaking, for much of the printed discourse one encounters, the literal level is the only level of meaning. That this is so will be evident at a moment's reflection. The author of a textbook, for instance, has as his main purpose the conveying of information as clearly and directly as possible. To achieve clarity and directness, the textbook author does not typically employ deception or indirection, purposeful ambiguity, a complex persona, irony, or other such literary and rhetorical techniques. His aim typically is not to call attention to his authorial virtuosity, but rather to be as straightforward and uncomplicated as he can. For this reason he will try to provide the kind of context which will most efficiently exclude unintended meanings and specify intended ones.

To reiterate, for much of what we read, the literal level is the only level of meaning, and apprehension of the literal level of meaning does not call into play complex, inferential, or other higher-level thought processes. To be sure, such processes do come into play as we reflect on what we read, but reflection is not a reading skill but a thinking skill.

The LCDT and the PRT were designed to measure literal comprehension. Both words in the term literal comprehension as used to describe what these two tests measure are significant. On the one hand, the word comprehension implies that the tests are not focusing on eye control, phonetic decoding, or other prerequisites of apprehension or understanding. On the other hand, the word literal indicates that the tests are measures of the explicit and clearly implied meanings of discourse, rather than of such additional meanings as require, for instance, complex or higher-level inferential,

analytic, synthetic, rhetorical, allusive, or critical reading or thinking skills.

The LCDT and the PRT, then, assume the possession of visual and phonetic skills prerequisite to comprehension but do not presume to assess those skills necessary for processing beyond the apprehension of the explicit meaning of the text.

Some examples may help to fix more clearly the limitations of the LCDT and the PRT as measures of literal comprehension. It is assumed, for instance, that the two tests measure (or indicate the possession of) the kind of skills or abilities required to apprehend the meaning of the following:

> As Mary skipped along the sidewalk, her shoelace came untied. She tripped and fell and bruised her knee.

Literal comprehension of these sentences would entail (1) understanding the grammatical or syntactic relations among the words, including noun/verb distinctions, verb inflections, and pronominalization; (2) apprehending the explicit meanings of the words, including what such words as Mary, skipped, sidewalk, shoelace, tripped, fell, bruised, and knee mean; and (3) understanding the clearly implied meanings that Mary bruised her knee <u>because</u> she fell and fell <u>because</u> she tripped and tripped <u>because</u> her shoelace came untied.

The skills involved in processing these two sentences are the skills of literal comprehension. It is clear that such skills do not include higher-level intellectual processing. While it is assumed that all reading comprehension involves inference at a low level (e.g., inferring that orthography contains potential meaning, that letters in sequence form words, that words symbolize sounds, and that particular sounds have particular mean-

ings in given contexts), it is also assumed that such inference as is required, say, for complex propositional logic is not involved in literal comprehension.

To take another example, in Keats's phrase,

"No, no, go not to Lethe . . .,"

the literal level involves grammatical and vocabulary knowledge, most of which would be possessed by fairly young school children (with perhaps the exception of the relatively unfamiliar Lethe). What the literal level does not involve is the kind of additional inferential processing which discovers (from context) that there is a speaker addressing someone who has sought a certain kind of advice. Also not involved in literal comprehension would be the perception of the urgency or emphasis of the first two words. Further, any allusions or suggestions called up by Lethe (the classical nether-world river of forgetfulness)--say, to Homer or to Greek mythology--would not be part of literal comprehension.

Similarly, given the following Swiftian product, "I am very sensible what a weakness and presumption it is to reason against the general humour and disposition of the world," literal comprehension would involve knowledge of the meanings of the words in this context. Literal comprehension would not involve the perception that this is a curious thing to say (i.e., apart from the possible unfamiliarity of some of the words), or the perception that the tone of the statement needs to be pursued and pinned down.

Literal comprehension, then, given the necessary visual abilities and orthographic-phonetic knowledge, requires possession of grammatical rules and semantic knowledge. To possess grammatical rules signifies the capacity to apply the principles which govern the positional and inflectional relationships among words. It does not necessarily entail the ability to state formulated rules explicitly. Thus, children can apply grammatical rules

without being able to express or define such rules.

The semantic capacity required for processing discourse involves two kinds of knowledge: (1) vocabulary (or "dictionary") knowledge, the ability to recognize a given word, and (2) a script or schema which permits the determination of the meaning of a given word in a given context. For instance, the word dogs, isolated from any surrounding or related context, has no determinate meaning. Only when dogs is used in a particular context does it take on a definite meaning. In the context of greyhound racing, dogs might have one meaning; used metaphorically by tired mail carriers, it might have another significance; the possibilities for varied semantic--and grammatical--significance clearly abound (e.g., "Hate dogs their flight . . ."). The semantic capacity properly to understand any word in a given context depends on a script which includes experience of that word in one of its particular contexts. If a person possesses grammatical rules, then literal comprehension requires only (1) that he have previously encountered and understood a given word and (2) that he have encountered and understood it in a context which delineated its meaning as the present context delineates it.

The LCDT and the PRT both appear to be measures of literal comprehension. That is, both tests seem to access the kind of grammatical and semantic knowledge and ability necessary for apprehending the explicit and clearly implied meanings of discourse without requiring higher-level skills. Since one of the principal purposes of this paper is to evaluate the practical potential of these two tests as measures of literal comprehension, the following sections will describe the construction of the tests. Then the paper will focus on the research studies which have involved the tests. The paper will conclude with evaluations of the performances of the LCDT

and the PRT, including practical recommendations concerning replicability, limitations, and modifications.

The Literal Comprehension Details Test (LCDT)

The LCDT was conceived as a battery of test passages, scaled by difficulty level, with accompanying rule-based items for measuring literal comprehension. The difficulty levels were interpreted from readability scores based on the Spache and Dale-Chall readability formulas. The passages were to be coherent and unified, and their lengths were to vary by difficulty level. The relationship between readability scores and difficulty levels, as well as approximate passage lengths, is illustrated in Table 1. Each passage was to

Table 1 about here

be accompanied by 2 main idea items, 2 title items, and 8 wh- detail items. Though some passages did not permit the achievement of this goal (e.g., it was not always possible to write 2 title items according to the formulated rules), on most passages the requisite number of items could be written. The completed corpus consisted of 300 passages, 15 at 20 different difficulty levels (spanning, approximately, grades 1 to 10), with usually 2 title items, 2 main idea items, and 8 wh- detail items. It should be noted at this point that since the rule-based title and main idea items were not used in the study to be reported, there will be no further discussion of them.

Passage Sources. Passages for the LCDT were from three different sources. Some were produced under contract for the Bureau of School and Cultural Research. Some were taken by Bureau staff from a variety of sources, and some were written by Bureau staff expressly for the LCDT. Upon reception and inspection, the passages produced under contract were found to require

Table 1

Readability Scores, Difficulty Levels, and Passage Lengths
Literal Comprehension Details Test

	Readability Scores	Difficulty Level	Passage Length (Words)	
S	1.0-1.4	1	30	
P	1.5-1.9	2	40	
A	2.0-2.4	3	50	
C	2.5-2.9	4	60	
H	3.0-3.4	5	70	
E	3.5-3.9	6	80	
	4.50-4.74	7	90	
	4.75-4.99	8	100	
D	5.00-5.24	9	110	
A	5.25-5.49	10	120	
L	5.50-5.74	11	130	
E	5.75-5.99	12	140	
	6.00-6.24	13	150	
	6.25-6.49	14	160	
C	6.50-6.74	15	170	
H	6.75-6.99	16	170	
A	7.00-7.24	17	170	
L	7.25-7.49	18	180	220
L	7.50-7.74	19	180	220
	7.75-7.99	20	180	220
			Expository	Narrative

significant editing and rewriting, both to attain desired prose standards of coherence and unity and to assure accurate readability scores. No passage received under contract escaped revision. The sources of these contractually-produced passages were encyclopedias, standardized test passages, and some textbooks. Bureau staff took many passages from such sources as literature anthologies, newspapers, magazines, and encyclopedias. Most of these passages required little or no editing, while some required moderate editing. The original passages written for the LCDT vary, as might be expected, from accounts of personal experience to academic-informative-general interest passages to rather fanciful pieces.

The exercise of accumulating a set number of passages of reasonably acceptable prose quality at given difficulty levels (i.e., within narrowly specified ranges of readability scores) requires considerable discipline. Most passages can be taken from the various sources or created without great vexation. Randomly-selected passages will span a great range of difficulty. But when, say, 75 percent of the passages have been collected at the required difficulty levels, or when the desired numbers of passages have been gathered for given difficulty levels, the exercise of producing (or locating) passages for specific difficulty levels can become rather grueling. It is largely for this reason that passages taken from printed sources required editing; if, for example, an extant passage had a difficulty level of 16 but the passages for level 16 had already been produced and more level 15 (or level 17) passages were required, then experimentation with easier (or harder) synonyms or with shorter (or longer) sentences had to occur. To alter passages without significantly distorting meaning and without producing barbarous prose placed great demands on sensitivity and concentration.

Item-writing rules. After the 300 readability-scaled passages were

completed, some experimentation occurred on items which would measure literal comprehension and which could be written according to rules. Both the rationale for rule-based items and the decision to write wh- detail items were based on the work of Bormuth (1970) on achievement test items. Essentially, Bormuth argues for rule-based items on the grounds that only such items avoid idiosyncracies and interpretations introduced by item writers; in a word, rule-based items give some measure of assurance of the avoidance of subjectivity or eccentricity.

Even casual inspection of reading comprehension items on standardized achievement tests reveals an enormous variety in the type and quality of items, even of items categorized by test-makers as having similar measurement properties. The question, of course, is whether obviously-different items can be measuring the same thing. Put another way, how can one interpret performance on such items?

Rule-based items represent a method of avoiding this uncomfortable question, for if the item-writing rules are clear, it is a simple matter to review the items for conformity. Items written acceptably to rules should be readily interpretable.

Bormuth's recommendation of wh- detail items [i.e., items with stems introduced by the following "wh" words: how, what (noun, pronoun), what (verb), when, where, which, who, why] follows centuries of standard pedagogical practice. Wh- detail questions are extremely useful for getting at the literal meaning of discourse. The writing of the wh- detail items for the LCDT departed in two ways from Bormuth's proposed methodology. First, because of time constraints, only one of each type of wh- detail item was to be written for each passage; thus, a maximum of eight wh- detail items could be written for each passage. Secondly, the rules for writing wh- detail items for the LCDT

permitted only verbatim items to be written. (Bormuth's illustrative items occasionally involve paraphrase or substitution of synonyms.)

For each LCDT passage, then, eight verbatim wh- detail items were to be written. Briefly, the procedure involved random selection of a sentence from a passage. Given a sentence, an attempt was made to write a verbatim wh- detail item. The eight types of wh- words were listed alphabetically, and an attempt was made to write the first wh- item type (i.e., "how") on the first sentence randomly selected. If the sentence did not permit a "how" item, an attempt was made to write a "what (noun, pronoun)" item. The item writer would try, for each new randomly-selected sentence, to write the next type of wh- detail item on the list. (If one was skipped, the item writer would return to it on the next sentence.) This procedure permitted the production of nearly eight verbatim wh- detail items for each of the 300 passages. The rules for writing wh- detail items for the LCDT are contained in Appendix A.

Each wh- detail item features a stem, introduced by the appropriate wh- word, and either 3 or 4 responses (3 for difficulty levels 1-4, 4 for difficulty levels 5-20). One response is the correct answer, and the other responses are both grammatically and semantically plausible. Dependent on the passage upon which it is based, a wh- item should not be answerable through application of test-wiseness skills. The distractors are taken verbatim from the passage whenever possible. This is another precaution introduced to assure that the passages be read before the items are answered. Traditional standards for objective items are also observed in the wh- detail items (e.g., avoidance of correct responses standing out because of greater length than distractors). All LCDT passages and items were thoroughly reviewed before tests were assembled from the passage-item battery.

The Paraphrase Reading Test

The PRT was developed for use as a criterion measure of literal comprehension. The occasion for the development of the PRT was a construct validation study of the Multiple-Choice Cloze (MCC) Exercises, 1725 modified cloze passages with accompanying multiple-choice items. The passages for the PRT are the same passages that are used in the MCC exercises. The PRT items are wh- detail items based on paraphrases of the sentences in the (MCC) passages. The basic difference between the PRT and LCDT items, then, is that the PRT items involve paraphrase.

The need for a construct validation study of the MCC exercises arose for several reasons. An earlier effort to validate the MCC, an effort which included the use of the LCDT, suffered from the lack of a standardized test of sufficient quality and interpretability. Another significant reason was that after the first validity study the MCC underwent substantial change, including reclozing of passages, replacement of many distractors, and removal of titles. Perhaps the most significant reason for the second study was a perceived theoretical shortcoming of the LCDT, which was used as a criterion measure in the initial study. It could be argued that items on the LCDT could be answered by application of such test-wiseness skills as orthographic or phonetic matching. In other words, it might be possible to answer LCDT items without reading the passages on which they are based. To the extent that such test-wiseness skills are employed in responding to the LCDT, the test is invalidated as a measure of literal comprehension.

Rationale. The use of paraphrase items was based on Anderson's (1972) defense of paraphrase as a valid measure of (literal) reading comprehension:

> The argument that paraphrase questions assess comprehension is very simple . . .. [I]n order to answer a question based on a paraphrase, a

> person has to have comprehended the original sentence, since a paraphrase is related to the original sentence with respect to meaning but unrelated with respect to the shape or sound of the words. (p. 150)

Further, the rules for paraphrase item writing were derived from Anderson's definition of paraphrase: "Two statements are defined as paraphrases of one another if 1) they have no substantive words (nouns, verbs, modifiers) in common and 2) they are equivalent in meaning" (p. 150). Paraphrase, then, was selected for use as the criterion measure in the second construct validation study of the MCC.

Passage sources. As stated, PRT passages are identical to those used in MCC exercises. The sources of MCC passages are textbooks and a wide variety of other printed materials, including newspapers, magazines, reference books, advertisements, and recipes. The passages are brief, never longer than 80 words and averaging 60 to 70 words. They are coherent, but they are too short to assure unity in the sense of a beginning, middle, and ending. The passages are taken as is from their sources, with no "correcting" of punctuation or grammar. Minor editing occurs very infrequently, and then only to assure coherence. "Clozed" passages average ten deletions or blanks, with a multiple-choice item for each blank.

For the PRT, only MCC passages taken from reading or literature texts were used. The deleted words were replaced in the blanks, and the passages were retyped. Then each sentence, clause, or long phrase was paraphrased, and wh- detail items were written for the paraphrases sentences, clauses, etc.

Paraphrase item writing. The rules for writing paraphrases were derived from Anderson's (1972) brief definition. However, paraphrases for the PRT were defined somewhat more restrictively than Anderson had required. For the PRT, synonyms or synonymous phrases used in paraphrasing were to come,

as far as possible, from among words at the same grade level as the passage containing the sentence to be paraphrased. To assure the grade level of the paraphrase vocabulary, graded word lists (Harris & Jacobson, 1972; Carroll, Davies, & Richman, 1971) were used, whenever possible. Some givens of paraphrasing included the impossibility of finding synonyms or synonymous phrases for proper nouns, auxiliary verbs, or the verb to be. MCC passages, the passages on which paraphrase items were to be written, are scaled using the same readability formulas as the LCDT, and experience quickly showed that it was not feasible to try to write paraphrases for the sentences in passages below difficulty level 7 (i.e., below grade 4, approximately). The problem with these is that many synonyms for words typically found in texts at such levels are too difficult; to use them would be to increase the difficulty and complexity of the task involved in responding to the paraphrase item. To do this would be in some measure to invalidate the items as measures of literal comprehension. Such items would place a heavier burden on verbal intelligence than the literal comprehension of the passage would require.

After each sentence, or significant part of a sentence, in a passage had been paraphrased, wh- detail items were written on the paraphrases. The rules (see Appendix B) for writing paraphrase items were adapted from the rules for writing wh-detail items developed for the LCDT. The basic difference between the item-writing procedures was that there was no restriction in the number or type of wh- items written for the PRT. For every paraphrase, all possible wh- detail items were written. The reason for this was that, as mentioned above, the passages were very short and as large a pool of items as possible was desired for each passage to facilitate test construction.

Though the intention was to control paraphrase vocabulary, to keep it

from exceeding the grade level of the source passage, it was very difficult, and occasionally not possible, to control paraphrase vocabulary on higher difficulty passages. Graded word sources were not adequate, which is the same as saying that synonyms for difficult words are often more difficult than the words for which they are to be substituted.

The basic rules for writing items based on the paraphrase differed little from the rules for writing wh- detail items for the LCDT. There was no requirement that distractors be taken from the passage, for example, but the greatest difference arose in response to the need to control for items which involved only partial paraphrases. In some cases, sentences, clauses, or phrases could not be completely paraphrased. That is, it was not always possible to find synonyms (paraphrases) for every content word. Usually, a substantial portion of a sentence could be paraphrased, so that there are no verbatim (unparaphrased) items, but in some cases either a stem or a response may be incompletely paraphrased. When a correct response was incompletely paraphrased, distractors were designed to prevent the successful exercise of such test-wiseness skills as orthographic matching.

There were 356 MCC exercises based on passages taken from reading or literature texts. The elimination of 122 (from grade 1-3 sources) left 234 passages. From these, 39 passages were selected randomly. Thus, 17 percent of the available passages were sampled for the construction of paraphrase items. (Some departure from randomness was necessitated because certain passages did not yield the requisite number of paraphrases.) An average of about a dozen paraphrase items was written for each passage, and the items were intensively reviewed both in-house and by reading professionals.

Construction and Administration of the LCDT

In the spring of 1975 the LCDT was administered to over 5,000 students in grades 1 through 9 in an upstate urban school district as part of a validity study of the MCC exercises. There were 36 forms of the MCC and 36 forms of the LCDT. The passages on the forms were never identical, and, except at the lower grade levels, seldom the same length, but the 36 MCC forms were parallel to the 36 LCDT forms. The forms for each test were divided into 3 levels, with 12 forms at a level. Students in grades 1-3 took Level I forms; students in grades 4-6 and 7-9 took Level II and Level III forms, respectively. The passages on each MCC form at a test level were parallel in difficulty to each other and to the passages on the LCDT forms for the same test level. Parallelism was controlled by readability scores (difficulty levels). On each test form, passages were arranged in ascending order of difficulty. The difficulty level ranges for the test levels of both the LCDT and the MCC are as follows:

	Test Level		
	I	II	III
Difficulty Level Range	1-10	5-16	11-20

For construction of the LCDT forms, pairs of difficulty levels were combined and their passages pooled in preparation for random selection of test passages. Thus, the first passage on each Level II test form was drawn randomly from the 30 available passages resulting from the pooling of the passages at difficulty levels 5 and 6. A similar procedure was followed for the selection of subsequent passages. The only variation from this method was at Level I, where difficulty levels 1 and 2 were discrete sampling units, and at Level III, where difficulty levels 11 and 12 were discrete sampling units.

For each LCDT passage on each form, five wh- detail items were chosen, for a total of 30 items per form. The items were selected randomly where feasible, but the overriding criteria for item selection were (1) avoidance of mutual cueing and (2) even distribution of wh- item types. Mutual cueing was defined as a stem of one item cueing the answer to another item. Even distribution of item types was achieved for all three test levels. In other words, there were not more "when" questions than "why" questions, for example, across the forms at a test level. A typical LCDT passage, with accompanying items, is illustrated in Figure 1.

The LCDT forms were administered one week after the administration of the MCC forms. Means and standard deviations and reliability and validity coefficients were calculated for all MCC and LCDT test forms. Also, data from Rasch analyses of the forms permitted some inspection of deviant items. In addition to the MCC and LCDT data, scores on the California Achievement Test (CAT) for students in grades 1-8 and scores on the Short Form Test of Academic Aptitude (SFTAA), an IQ measure, for students in grades 1-6 were obtained. The CAT and SFTAA scores were entered into the validity correlation analyses (O'Reilly, Schuder, & Kidder, 1975), and the SFTAA data permitted some factor analyses (O'Reilly & Streeter, 1977).

As shown in Table 2, means and standard deviations for the LCDT (and for the MCC) were quite consistent, thus suggesting that parallelism among forms at a test level had been achieved and that the rules for writing wh-detail items had been applied with a high degree of consistency. Kuder-Richardson Formula 20 reliability coefficients for the LCDT and the MCC are reported in Table 3. As illustrated, the average K-R 20 for both tests is high, indicating again the consistency of both measures. Validity correlations for the LCDT and the MCC are given in Table 4. At test levels I and II,

During World War II, Britain was defended by an heroic air force, but it was difficult to keep the planes aloft. Fuel and spare parts were hard to get, but the worst problem was the fog which usually covered the airfields.

London is known especially for its dense fog. Since the city is near the ocean, the moist air seeps over the city and its airports, cools, and changes to fog. Before pollution control, smoke from homes and factories stuck to the fog which took on the yellow-green of pea soup. This green fog made it dangerous for planes to take off or land.

To keep their war planes flying, the English developed a method for clearing the fog from the airports. They lighted oil burners along the runways, and warm air rushing upwards carried the fog with it to 2,000 feet or more. Planes could then fly and carry on the defense of Britain.

21. Where is the city of London?

A. near the ocean
B. on seven hills
C. in Europe
D. on a wide river

22. What kind of fog made it dangerous for planes to take off or land?

A. white
B. green
C. grey
D. dirty

Figure 1. Sample Literal Comprehension Details Test Passage with Accompanying Items.

23. Who developed a method for clearing the fog from the airports?

A. the English
B. the Irish
C. the French
D. the Dutch

24. When was Britain defended by an heroic air force?

A. after the fall of Paris
B. after the attack on Normandy
C. during World War I
D. during World War II

25. Why was it difficult to keep the planes aloft?

A. because many pilots had been killed
B. because bombs were falling on the airfields
C. because of the fog which usually covered the airfields
D. because London is near the ocean

Figure 1 (Cont.) Sample Literal Comprehension Details Test Passage with Accompanying Items.

the correlations are moderately high to high; such correlations, demonstrating the high percentage of shared variance between the two measures, give strong support to the conclusion that both tests are measuring the same thing (i.e., literal comprehension). It may be noted here that factor analyses (reported in O'Reilly and Streeter, 1977) resulted in two factors, which were interpreted as a literal comprehension factor and an IQ factor. The MCC and the LCDT loaded heavily on the literal comprehension factor.

Tables 2, 3, & 4 about here

As part of the analysis of the LCDT, an attempt was made to identify and study the causes of item deviance. To date the analysis is incomplete, but preliminary efforts attempted to identify possibly deviant items by means of z-scores. (The z-scores were calculated for the items on each passage, using average percent correct of the items on a passage and the standard deviation of the passage items.) Negative z-scores lower than approximately -1.2 identified apparently or statistically deviant items. Perhaps 15 percent of the items on the LCDT forms were thus identified. Inspection of these items, however, frustrated in many cases attempts to explain their apparent deviance. Some items were clearly and explainably deviant. For example, extreme awkwardness of item stems and competitive (i.e., arguably correct) distractors were among the reasons given to account for actual deviance. As stated above, this phase of the analysis is not yet complete. It is expected that the completed analysis of LCDT item deviance will yield generalizations concerning the proportion of explainable deviant items and the relationship between z-scores and explainable deviance.

Table 2

Means and Standard Deviations for the MCC and the LCDT

Level	MCC Form	N	$\bar{X}$	S.D.	LCDT Form	N	$\bar{X}$	S.D.
I (Grades 1, 2, 3)	1	128	21.03	10.55	37	127	18.80	7.15
	2	126	20.26	10.49	38	124	19.52	7.32
	3	130	21.51	10.59	39	126	19.57	7.29
	4	124	22.44	11.34	40	121	19.02	7.30
	5	126	23.06	11.24	41	119	18 43	7.72
	6	126	19.71	9.24	42	122	19.17	7.00
	7	127	21.47	11.22	43	124	19.05	7.28
	8	127	18.84	10.40	44	124	19.20	7.53
	9	129	21.98	11.31	45	131	19.10	7.69
	10	127	20.47	10.43	46	123	19.19	7.47
	11	120	23.39	11.25	47	121	18.65	7.18
	12	123	22.67	11.41	48	121	20.12	7.69
II (Grades 4, 5, 6)	13	147	41.46	11.45	49	147	22.74	5.57
	14	151	40.01	14.11	50	153	21.95	5.46
	15	153	38.73	12.51	51	148	22.72	4.85
	16	152	40.99	11.62	52	152	22.74	5.66
	17	146	42.18	12.60	53	145	23.52	5.46
	18	151	36.35	11.03	54	144	23.19	5.45
	19	152	41.80	13.48	55	145	22.96	5.00
	20	148	42.00	12.08	56	149	22.60	5.96
	21	152	41.39	11.37	57	147	20.76	6.11
	22	152	39.63	13.57	58	148	22.19	5.86
	23	148	41.72	12.99	59	157	23.76	5.51
	24	149	39.01	13.32	60	145	21.87	5.73
III (Grades 7, 8, 9)	25	167	36.60	12.53	61	163	23.81	5.54
	26	164	36.44	11.69	62	162	23.89	7.01
	27	160	38.86	14.33	63	164	24.25	5.79
	28	161	40.47	12.82	64	161	23.89	4.83
	29	158	39.17	11.52	65	165	23.53	4.75
	30	165	42.54	13.35	66	166	21.20	6.20
	31	158	39.46	12.45	67	154	24.88	4.85
	32	163	37.07	12.01	68	163	22.40	5.65
	33	166	37.38	11.98	69	164	24.02	4.99
	34	159	38.08	13.60	70	156	22.01	5.31
	35	163	37.82	13.18	71	163	23.16	5.94
	36	165	41.82	12.56	72	154	22.03	6.90

Table 3

Kuder-Richardson Formula 20 Reliability Coefficients for the MCC and the LCDT

	MCC					LCDT				
Level	Form	N	I	KR-20	SE	Form	N	I	KR-20	SE
I (Grades 1,2,3)	1	128	41	.94	1.73	37	127	30	.92	2.02
	2	126	41	.95	1.64	38	124	30	.94	1.79
	3	130	41	.96	1.46	39	126	30	.90	2.30
	4	124	41	.96	1.46	40	121	30	.90	2.31
	5	126	41	.95	1.73	41	119	30	.91	2.32
	6	126	39	.95	1.57	42	122	30	.91	2.10
	7	127	41	.96	1.45	43	124	30	.93	1.92
	8	127	39	.96	1.51	44	124	30	.90	2.38
	9	129	41	.97	1.33	45	131	30	.90	2.43
	10	127	41	.96	1.49	46	123	30	.91	2.24
	11	120	41	.96	1.43	47	121	30	.94	1.76
	12	123	41	.96	1.54	48	121	30	.92	2.18
	Median			.96	1.49				.91	2.21
II (Grades 4,5,6)	13	147	60	.97	1.98	49	147	30	.93	1.47
	14	152	60	.96	2.82	50	153	30	.93	1.44
	15	153	60	.96	2.50	51	148	30	.90	1.53
	16	152	60	.96	2.32	52	152	30	.86	2.11
	17	146	60	.97	2.18	53	145	30	.93	1.44
	18	151	60	.94	2.69	54	144	30	.92	1.54
	19	152	60	.97	2.33	55	145	30	.85	1.94
	20	148	60	.95	2.69	56	149	30	.95	1.33
	21	152	60	.95	2.53	57	147	30	.94	1.50
	22	152	60	.97	2.35	58	148	30	.91	1.76
	23	148	60	.97	2.25	59	157	30	.94	1.35
	24	149	60	.95	2.97	60	147	30	.93	1.51
	Median			.96	2.35				.93	
III (Grades 7,8,9)	25	167	60	.96	2.51	61	163	30	.91	1.66
	26	164	60	.95	2.61	62	162	30	.94	1.71
	27	160	60	.96	2.87	63	164	30	.96	1.16
	28	161	60	.97	2.22	64	161	30	.89	1.60
	29	158	60	.96	2.30	65	165	30	.89	1.57
	30	165	60	.97	2.31	66	166	30	.94	1.52
	31	158	60	.95	2.78	67	154	30	.96	0.97
	32	163	60	.96	2.40	68	163	30	.90	1.78
	33	166	60	.95	2.67	69	164	30	.96	0.99
	34	159	60	.95	3.03	70	156	30	.93	1.40
	35	163	60	.97	2.28	71	163	30	.95	1.32
	36	165	60	.97	2.17	72	154	30	.95	1.56
	Median			.96	2.40				.94	1.54
Overall	Median			.96					.92	
	Mean			.96					.92	
	Range			.94-.97					.85-.96	

Note. N = number of subjects.
I = number of items.

Table 4

Zero-Order Correlations of MCC Scores with LCDT Scores

Test Level		
I	II	III[a]
.81	.73	.62

[a]Level III correlations do not include grade 9 data.

Construction and Administration of the PRT

The PRT was designed, as noted previously, as a literal comprehension criterion measure for use in a construct validity study of the MCC. In the spring of 1977, the PRT, the MCC, and three other measures were administered to students in grades 3, 6, and 9 in one metropolitan New York district and two upstate districts, one urban and one suburban. The schools and classes in the schools were selected for their socioeconomic and academic representativeness. The three other measures were the Gates-MacGinitie Reading Tests--Comprehension (Gates), the Stanford Achievement Test--Reading Comprehension (SAT), and the Degrees of Reading Power Test (DRP), currently under development in the New York State Education Department. Intercorrelational results of the MCC with all the four other measures may be obtained on request from the Bureau of School and Cultural Research. For purposes of this paper, only results involving the PRT, MCC, and Gates will be reported.

Approximately 1,350 students received either the PRT and the MCC, the PRT and the Gates, or the MCC and the Gates. The tests were administered one week apart. The actual test combinations are listed below:

Grade		
3	6	9
PRT/MCC	PRT/MCC	PRT/MCC
PRT/Gates Primary C	PRT/Gates Survey D	PRT/Gates Survey E
MCC/Gates Primary C	MCC/Gates Survey D	MCC/Gates Survey E

The Gates was used in this construct validation study because of its reputation as principally a measure of literal comprehension. High correlations were expected among the three measures; if such correlations were obtained, they would be interpreted as constituting strong evidence for the validation of the MCC as a measure of literal comprehension. Similarly, high correlations

would also validate the PRT as a measure of literal comprehension; the PRT, of course, has greater face validity than the MCC as a measure of literal comprehension.

There were three test levels for the PRT and the MCC, and three parallel test forms were constructed at each level. That is, the passages on the forms at each test level shared the same range of difficulty, and increases in difficulty from passage to passage were identical. Each PRT test form had five passages, and there were six items for every passage. The six items, selected from the pool of items written for each passage, were chosen on the basis of two criteria: (1) avoidance of mutual cueing and (2) quality (e.g., avoidance of awkwardness and ungrammaticalness). A PRT passage, with its items, is illustrated in Figure 2.

Means and standard deviations, Kuder-Richardson Formula 20 reliability coefficients, and Pearson Product-Moment correlation coefficients were computed for the three tests and are reported, respectively, in Tables 5, 6, and 7.

The means and standard deviations for the PRT suggest a good deal of consistency across the test forms, which in turn implies a degree of success in applying the item-writing rules and in attaining parallelism among test forms. (For future reference the relatively low standard deviations for the PRT forms at grade 9 should be noted here.) The very high K-R 20's for both the PRT and the MCC are evidence of the internal consistency of both measures and of the consistency of student responses to the PRT and MCC formats. The correlation coefficients are also high, as expected, especially at grades 6 and 3. At grade 6 the correlations indicate that approximately 70 percent of the shared variance for the three test combinations is accounted for by the same trait, i.e., literal comprehension. These correlations give strong support to the contention that all three tests are measures of literal comprehension.

The two gods took on the appearance of poor wayfarers and wandered through the land, knocking at each lowly hut or great house they came to and asking for food and a place to rest in. Not one would admit them; every time they were dismissed insolently and the door barred against them. They made trial of hundreds; all treated them in the same way.

(63) What did the two divine beings do?

1. pretended to be lost and sick
2. assumed the likeness of needy travelers
3. acted like common gentlemen
4. appeared as great and worthy citizens

(64) Where did the two divine beings roam?

1. throughout the country
2. throughout their palaces
3. everywhere but the market place
4. only in the forest

(65) What did the two divine beings request?

1. somewhere to wash and rooms to sleep in during the night
2. water to drink and a place to clean up in
3. a place to pray and some water to drink
4. something to eat and a spot to pause and relax in

(66) What would nobody do?

1. turn the two gods away
2. let the two gods in
3. admit that they had room and food to spare
4. permit the two gods to leave

Figure 2. Sample Paraphrase Reading Test Passage with Accompanying Items.

(67) What did the two divine beings do at every humble shanty or fine mansion they arrived at?

1. rang the doorbell
2. rapped on the door
3. stared in the window
4. stood by the gate

(68) How did everyone behave toward the gods?

1. courteously
2. similarly
3. pleasantly
4. differently

Figure 2 (Cont.) Sample Paraphrase Reading Test Passage with Accompanying Items.

Table 5

Means and Standard Deviations for PRT, MCC, and Gates
by Grade Level and Test Combination Group

PRT				MCC				Gates		
Form	MCC Group	Gates Group	Combined	Form	PRT Group	Gates Group	Combined	PRT Group	MCC Group	Combined
					Grade 3					
321	19.5(6.0) N=75	20.0(6.3) N=61	19.8(6.2) N=136	311	36.8(10.0) N=75	34.2(11.6) N=73	35.5(10.8) N=148	33.3(9.5) N=184	29.7(12.3) N=227	31.5(10.9) N=411
322	18.3(6.8) N=81	19.5(7.3) N=62	18.9(7.1) N=143	312	33.2(11.2) N=81	30.4(12.4) N=79	31.8(11.8) N=160			
323	17.6(6.4) N=80	18.0(5.7) N=61	17.8(6.1) N=141	313	33.7(11.4) N=80	31.4(11.9) N=75	32.6(11.7) N=155			
					Grade 6					
621	22.2(6.8) N=81	23.1(4.7) N=50	22.6(5.8) N=131	611	40.3(10.3) N=81	39.2(9.3) N=75	39.8(9.8) N=156	40.7(8.6) N=174	39.6(9.3) N=227	40.2(9.0) N=401
622	19.4(5.5) N=84	18.9(5.7) N=55	19.2(5.6) N=139	612	36.1(9.9) N=84	33.0(10.0) N=77	34.6(9.9) N=161			
623	18.8(7.0) N=76	19.6(6.2) N=59	19.2(6.6) N=135	613	35.0(10.6) N=76	35.1(8.8) N=75	35.1(9.7) N=151			
					Grade 9					
921	18.9(4.3) N=87	20.6(4.2) N=71	19.8(4.3) N=158	911	40.2(8.0) N=76	40.3(8.5) N=69	40.2(8.3) N=145	39.5(10.2) N=227	38.0(10.5) N=210	38.8(10.4) N=437
922	19.4(5.0) N=69	19.6(4.9) N=75	19.5(5.0) N=144	912	41.4(9.8) N=69	41.4(8.5) N=69	41.4(9.2) N=138			
923	17.4(4.7) N=72	19.5(5.2) N=81	18.5(5.0) N=153	913	36.0(9.1) N=72	36.0(8.0) N=72	36.0(8.6) N=144			

Table 6

Kuder-Richardson Formula 20 Reliability Coefficients for the Multiple-Choice Cloze Test and Paraphrase Reading Test by Grade Level

Multiple-Choice Cloze			Paraphrase Reading Test		
Form	N	K-R-20	Form	N	K-R-20
		Grade 3			
311	321	.97	321	153	.96
312	329	.95	322	157	.96
313	314	.96	323	152	.90
Average K-R-20 = .96			Average K-R-20 = .94		
		Grade 6			
611	305	.97	621	152	.97
612	320	.96	622	155	.90
613	299	.94	623	146	.95
Average K-R-20 = .96			Average K-R-20 = .94		
		Grade 9			
911	336	.98	921	173	.94
912	332	.97	922	171	.91
913	336	.95	923	182	.92
Average K-R-20 = .97			Average K-R-20 = .92		

Table 7

Zero-Order Correlations

Grade	PRT-MCC	PRT-Gates	MCC-Gates
3	.80	.79	.76
6	.84	.83	.84
9	.68	.48	.76

Table 7

Zero-Order Correlations

Grade	PRT-MCC	PRT-Gates	MCC-Gates
3	.80	.79	.76
6	.84	.83	.84
9	.68	.48	.76

As with the LCDT, part of the analysis of the PRT results involved examination of deviant test items. Deviant items were tentatively identified by means of z-scores, and the items so identified were inspected for the sake of determining the causes of actual deviance. Inspection of the PRT items is as yet incomplete, but preliminary findings indicate that PRT items are deviant in slightly higher proportions than are LCDT items. Discoverable causes of PRT item deviance seem closely related to problems involved in making paraphrases. Further study will attempt to determine the relationship between statistical deviance and explainable (actual) deviance.

Discussion

LCDT

The LCDT has high face validity as a measure of literal comprehension. Its items require no propositional inference, no drawing of conclusions, no analysis or synthesis of ideas. The test data confirm the consistency of the measure and of the application of the rules for writing wh- detail items. These two points, in concert with moderately high correlation coefficients and factor analytic findings, provide fairly strong grounds for the validation of the LCDT as a measure of literal comprehension. Certainly the LCDT battery of passages and items represents a resource of high potential.

The LCDT does have one possible shortcoming. Because its items are verbatim items, test-wise students may answer them correctly without reading

or comprehending the passage by a process of orthographic or phonetic matching. (The slightly transformed stems and correct responses may be located in the passage.) It is for future study to determine the extent to which such test-wiseness techniques are employed under actual test-taking conditions. It would seem unlikely that test-wiseness would come into play frequently enough to invalidate results for a single test administration. The practical question is whether test-wiseness could increasingly become a factor across several test administrations, say, in an achievement monitoring design.

PRT

Face validity for the PRT is higher than for the LCDT. There is no problem of orthographic-phonetic matching with PRT items. The consistency of the PRT and of the application of its item-writing rules is attested to in the data. These factors, combined with high validity coefficients, provide very strong support for the PRT (and also for the MCC and the Gates) as a measure of literal comprehension.

The relatively low correlation between the PRT and the Gates for grade 9 warrants comment. Part of the explanation is statistical. The distribution of Gates scores for grade 9 are positively skewed and the variability of scores on the PRT is somewhat less for grade 9 than for grades 3 and 6. These two factors partially explain the relatively low correlation. Much of the correlation is explicable in terms of shortcomings of the PRT forms for grade 9, however.

Two problems occurred in writing the paraphrase items for the grade 9 forms; the problems did not occur exclusively at the grade 9 level, but they were more pervasive at that level. One problem involved the writing of paraphrases and the other involved the increasing length of item stems and responses.

An effort was made to control paraphrase vocabulary so that it did not exceed the grade level of the passage source. This could be done fairly consistently on passages at grade 6 and below; available graded word lists facilitated vocabulary control for the paraphrasing of these passages. But for passages taken from sources above grade 6, available graded word lists were inadequate as sources of synonyms. Words which would serve as acceptable synonyms did not appear on the graded word lists. Thus, paraphrase vocabulary increased in difficulty on passages above grade 6, and the proportion of such passages was much higher on the grade 9 forms.

The second problem, increasing length of item stems and responses, was a function of the more difficult passages which appeared on the grade 9 forms. By definition, more difficult passages feature higher proportions of long sentences. Paraphrases of long sentences will themselves be long. And greater stem and response length contributes to greater item difficulty.

In other words, application of the paraphrase technology in producing items for the grade 9 forms elevated the difficulty of the items on those forms. One further piece of evidence illustrating the problem with the grade 9 forms lies in the relationship between PRT and MCC test forms. For grade 9 the PRT forms were relatively much more difficult in comparison to the MCC forms than they were at grades 3 and 6. This additional evidence further confirms the increased difficulty of the grade 9 PRT forms. The relatively low grade 9 correlations between the PRT and the Gates (and even the somewhat lower correlations between the PRT and the MCC at grade 9), then, can be largely understood as the result of problems in the application of the paraphrasing and item-writing rules.

Conclusions

The findings of this investigation into the feasibility of producing rule-based measures of literal comprehension are very positive. Application of the rules developed for writing both verbatim wh- detail items and paraphrase items was successful. The rule-based items permitted construction of test forms with high degrees of consistency and reliability and strong evidence of validity.

Neither the LCDT nor the PRT was without problems, however. The verbatim wh- detail items of the LCDT are open to the charge that they can be answered by the application of test-wiseness skills. No obvious solution to this problem comes immediately to mind. As stated above, further research might profitably investigate the extent to which such test-taking skills contaminate test results. Also, future investigation could be applied to the solution of the test-wiseness problem.

The problem with the PRT, that the items became disproportionately difficult on the upper-grade-level passages, is not insoluble. In fact, the problem is at least as much attributable to the constraints upon item-writing imposed by the brief passages used on the PRT forms as it is to the paraphrase item technology. The obvious solution to the problem is to write paraphrase items on longer passages; for example, the passages on the LCDT. Longer passages would permit much greater flexibility in the writing of paraphrase items because they would contain more sentences for which acceptable paraphrases could be written. With the short passages used on the PRT, paraphrases had to be forced for the sake of accumulating six items per passage. With longer passages and more flexibility in test construction, poor quality paraphrases would no longer have to be written.

Whether the use of longer passages would permit the extension of the paraphrase technology to passages from sources below grade four is conjec-

tural. It would seem that the paraphrase writing rules could be applied to longer passages even at such low grade levels. Lengthened passages might also alleviate the problem of controlling for vocabulary difficulty at upper grade levels.

Several practical recommendations arise from this analysis of the LCDT and the PRT. The first recommendation is that the LCDT be used; it is an extant resource which could serve in achievement monitoring designs, for example, or it could be used instructionally if teachers had it to use. A corollary of this recommendation is that the range of the passages, presently 20 difficulty levels, be extended at least to 26 difficulty levels to increase the test's utility for upper-grade students, many of whom would quickly top out on the extant passages.

Another recommendation is that given the length of the LCDT passages, they would be very suitable to the application of paraphrase item technology. Paraphrase items should be written for LCDT passages, then; if this suggestion were followed, all possible paraphrase items should be written on each passage. Such items, with their superior face validity, would constitute an extremely valuable resource for the measurement of literal comprehension.

The original design of the LCDT called for a maximum of eight wh- detail items per passage. It is here recommended that the number of items be increased by the writing of all possible verbatim wh- detail items on each LCDT passage. (The task is a finite one if the items are verbatim.) The larger pool of items resulting from this exercise would greatly increase the flexibility and utility of the resource.

It is clearly more difficult to write paraphrase items than it is to write verbatim wh- detail items. The rewards are greater, though, and this should be kept in mind if such options are ever seriously considered.

One final remark. In the application of the paraphrase item-writing rules, care should be taken to avoid forcing paraphrases where no adequate ones present themselves. Forcing could result in either unconscionably awkward or barbarous-sounding paraphrases or paraphrases which grow increasingly metaphorical. Either excess has an invalidating effect on the paraphrase item as a measure of literal comprehension. Judgment and sensitivity, then, must be exercised in the application of item-writing rules (and in the review and selection of items for test form construction).

There is much to be said for rule-based approaches to the measurement of reading comprehension, but one must be wary of the temptation to assume that reading comprehension measures can be completely automated or mechanized. Labor under such delusion must surely conclude in frustration.

Members of the research community interested in pursuing these suggestions may have access to the materials already prepared and make use of the rules accompanying this paper.

References

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42 (2), 145-170.

Bormuth, J. On the theory of achievement test items. With an appendix by P. Menzel: On the linguistic bases of the theory of writing items. Chicago: University of Chicago Press, 1970.

Carroll, J., Davies, P., & Richman, B. The American Heritage word frequency book. New York: Houghton-Mifflin, 1971.

Construct Validation of Multiple-Choice Cloze Exercises. Report in preparation, Bureau of School and Cultural Research, New York State Education Department, Albany, NY.

Dale, E., & Chall, J. A formula for predicting readability: Instructions. Educational Research Bulletin, 1948, 27, 37-47.

Harris, A., & Jacobson, M. Basic elementary reading vocabularies. New York: Macmillan, 1972.

Kidder, S. J. Apprehending text-based meaning from a Piagetian perspective. Paper presented at the annual meeting of the New York State Reading Association, Pre-Conference Institute #4 entitled Piagetian Theory and Reading, Concord Hotel, Kiamesha Lake, New York, November 1976.

O'Reilly, R. P., Schuder, R. T., & Kidder, S. J. Validation of a multiple-choice cloze test of literal comprehension: Summary report. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

O'Reilly, R. P., & Streeter, R. E. Report on the development and validation of a system for measuring literal comprehension in a multiple-choice cloze format: Preliminary factor analytic results. Journal of Reading Behavior, 1977, 9 (1), 45-69.

Schuder, R. T., Kidder, S. J., & O'Reilly, R. P. Defining and measuring the literal comprehension of written discourse. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Spache, G. A new readability formula for primary-grade reading materials. Elementary School Journal, 1953, 53, 410-413.

Spache, G. Good reading for poor readers. Champaign, Ill.: Garrard Press, 1960.

APPENDIX A

RULES FOR CONSTRUCTING WH- DETAIL ITEMS

WH- Detail Items

Format: Levels 1-4, 3 responses
Levels 5-20, 4 responses

1. Given a passage:

2. Randomly take a sentence number from a permutation block representing all possible sentences in the passage (in this case, 1-16).

   2.1. Take numbers from left to right across the block and so on down through the entire block if necessary; if block is exhausted before the passage, use next block; always start a passage with a new block.

   2.2. If number taken from block does not represent a sentence in the passage (e.g., 15 when there are only 10 sentences), take the next number.

3. Starting at the top, take a detail question type from the following alphabetical list (see attachment for illustrative examples of detail question types):

   HOW
   WHAT--noun, pronoun
   WHAT--verb
   WHEN
   WHERE
   WHICH
   WHO(M)
   WHY

4. If possible, write the detail question about the sentence taken in I. 2.

   4.1. Write clear, concise questions in colloquial English, changing the wording of the sentence as little as possible. (Exception: replace pronouns with their referents.)

   4.1.a. Begin each question with the appropriate detail word (e.g., how, what, etc.).

   4.2. Avoid anaphora when possible.[1]

   4.3. Avoid inference.[2]

   4.4. Ask each detail question only once per passage.

   4.5. If possible, ask all 8 detail questions of each passage.

4.6. Ask only one detail question per sentence unless the sentence or passage is rich in detail and there are few sentences, in which case repeat I. 2. from a new permutation block until all 8 wh-questions have been asked if possible.

5. If the detail question cannot be asked of the sentence taken in I. 2. (e.g., there is no answer to a "how" question), go on to the next detail question until a detail question is asked of the sentence if possible.

5.1. If a detail question cannot be asked of a given sentence, return to that same detail question first on the next sentence taken (e.g., if "how" is skipped, return to "how" first on the next sentence).

6. Take the next sentence number in the permutation block and ask the next detail question until all the detail questions are exhausted if possible (Some passages may not be rich enough in detail to provide bases for all eight detail question types.).

7. If possible, take the distractors from the passage verbatim.

7.1. Write only grammatically and semantically plausible distractors.

7.2. Write parallel distractors when possible.

7.3. Write distractors that closely match the correct response in number of words.

7.4. If distractors are not parallel or equal in length, write at least one distractor that parallels or matches in length the correct response.

7.5. Write no distractors that could be correct in the context of the passage.

7.6. Write distractors that are appropriate to the level of the passage.

8. If distractors cannot be taken verbatim from the passage,

8.1. Take distractors from the passage, changing them as little as possible in order to make them parallel and grammatically and semantically plausible (e.g., add determiners, adverbs, subordinators, etc.; or change verb tense, number, etc.; delete words; join words from scattered places in the passage).

8.2. If parallel, plausible distractors cannot be found in the passage, or if such distractors make the correct response debatable, take distractors from outside the passage. Such distractors must meet all the criteria in I. 7.1. to I. 7.6. above.

Footnotes

[1]The referent for a pronoun may be in preceding sentences. Adverbs like "soon" or "then" may refer to actions or situations in preceding sentences.

[2]The only exceptions would be passages where the logical relationship between two or more sentences is clearly implied. For example: "Carmen is writing to her friend, Carlos. Next Saturday will be his birthday." Why is Carmen writing to Carlos? Because next Saturday will be his birthday. Because is not in the passage but is logically and clearly implied as an expression of the relationship between the two sentences, "Tim, the turtle, has a new shell. He is very happy." Why is Tim happy? Because he has a new shell.

Illustrative WH- Detail Items

Wh-	Type	Example Q.	Example A.
How	Adverbial	Q. How many...?	A. 30, 40, etc.
		Q. How tall was the tree?	A. very tall
		Q. How are shoes made?	A. with leather
		Q. How did the brook flow?	A. rapidly
	Verb	Q. How does John get to school?	A. drives
	Adjectival	Q. How did Mary look?	A. sad, happy, pretty, etc.
What	Noun, Pronoun	Q. What did Jim need?	A. help
		Q. What did John eat?	A. lunch, ice, cream, it
		Q. What swam fast?	A. the fish
What	Verb	Q. What did Tim do?	A. ran, ate, slept, fell, etc.
		Q. What does Jane do?	A. sings, laughs, etc.
		Q. What was Harry doing?	A. thinking, talking, etc.
When	Adverbial-result	Q. When did the popcorn pop?	A. when the steam inside expanded
	Adverbial-time	Q. When did the boys come home?	A. in the evening, after school, at 4 o'clock, etc.
Where	Adverbial	Q. Where did Jack go?	A. for a walk, outside, to town, to New York
Which	Adjectival	Q. Whose cat was it?	A. Tom's, Mary's, John's
		Q. Which hat did Davy wear?	A. coonskin, blue, floppy, big
		Q. What kind of outfit did he wear?	A. new, old, dirty
		Q. What color was Bill's shirt?	A. blue, red, white
Who	Noun, person name (or pronoun standing for person)	Q. Who played ball?	A. Herbie, the boys, the players, he, they, etc.
		Q. Whom did the car hit?	A. Herbie, them, him, her, Mary, etc.
Why	Adverbial-cause, explicit	Q. Why did Tom trip?	A. because his shoes were too big
	Implicit	Q. Why did the ice melt?	A. The sun got very hot.

APPENDIX B

RULES FOR CONSTRUCTING PARAPHRASE

ITEMS FOR PAM ACHIEVEMENT MONITORS

I. Passage Selection

A. Determine range of difficulty for test forms.

1. Identify each difficulty level in the Reading/Literature MCC Exercises from which passages will be drawn.

2. Draw randomly the requisite number of exercises at each difficulty level.

3. Replace deleted words in blanks in each MCC exercise drawn.

II. Paraphrasing Selected Exercise Passage *

A. Number each sentence in every exercise passage.

1. In passages with compound sentences, number each main clause.

2. In passages with complex sentences, number each main clause, subordinate clause, and long modifying phrase.[1]

B. Paraphrase[2] each numbered sentence or clause.

1. If possible, replace all substantive words (nouns, verbs, modifiers[3]) with synonyms[4] (i.e., equivalent words or phrases).

a. Consult when necessary a dictionary, thesaurus, or dictionary of synonyms.

b. Consult other relevant reference words as necessary.

2. Proper nouns and pronouns often cannot be paraphrased.

3. Auxiliary verbs and the verb <u>to be</u> cannot always be paraphrased.

4. If possible, paraphrase vocabulary should not exceed the vocabulary level of the passage (as determined by difficulty level).

a. Consult Harris and Jacobson, 1972, when necessary.

b. Consult Carroll, Davies, and Richman, 1971, when necessary.

5. Retain meaning of original sentence (i.e., vocabulary and syntax of paraphrase should not involve significant alteration of the literal meaning of the original sentence).

---

* Rules for paraphrasing are based on Anderson's (1972) definition of paraphrase.

C. Flexibility in the writing of paraphrases is illustrated below:

1. A paraphrase does not have to have the exact number of words as the original sentence; it may be slightly longer or shorter.

2. Syntax may be altered in various ways.

   a. Order of clauses or phrases may be changed as long as literal meaning is retained.

   b. Voice of verbs may be changed (e.g., active to passive).

   c. Phrases may replace single words (and vice versa).

III. Writing Items for Paraphrased Passages[5]

A. Write WH-detail items on each paraphrased sentence, clause, or phrase. Adhere as much as possible to the following rules:

1. Write clear, concise questions in colloquial English, changing the wording of the paraphrase as little as possible. (Exception: replace pronouns with their referents.)

2. Begin each question with the appropriate detail word (e.g., how, what, when, where, etc.).

3. Avoid writing inferential WH-detail items (e.g., do not write a "why" item unless the causal relationship is either explicit or clearly implied in the text).

4. Write as many WH-detail items as possible for each paraphrase.

5. Try to write as least two WH-detail items for each paraphrase Note: Requirement for test forms was six WH-detail items/ passage. Passages are very short (50-80 words).[6]

B. Write three distractors for each item (i.e., four responses, including distractors and correct response).

1. Write only grammatically and semantically plausible distractors.

2. Write parallel distractors when possible.

3. Write distractors that closely match the correct response in number of words.

4. Avoid writing response arrays in which the correct response characteristically stands out because of its brevity, length, or syntax.

5. Write no distractors that could be correct in the context of the passage.

6. Write distractors that are appropriate to the difficulty level of the passage (see II. B. 4, above).

IV. Problems and Responses

A. Paraphrases

1. Not every sentence yields an adequate paraphrase, For example, vocabulary levels, uniqueness of vocabulary or structure, and other factors may make paraphrasing difficult.

2. When sentences which cannot be acceptably paraphrased result in passages which do not yield the requisite number of items, select another passage randomly from the relevant difficulty level.[7]

B. Items

1. When item stems contain substantive words verbatim from the passage, make sure correct response is not verbatim (i.e., do not write verbatim WH- detail items).

2. When a correct response is verbatim, make sure that some distractors are also verbatim to diminish the possibility of orthographic matching.

3. When a correct response is partially verbatim (e.g., this occurs occasionally in longer responses), make sure at least one distractor contains the verbatim element which appears in the correct response (to diminish orthographic matching).

Footnotes

[1]Extracted from context, subordinate clauses and some phrases may be paraphrased as main clauses or sentences. Example: "But even [a liar's invention] , being an empty thing that offers no hold . . ." is paraphrased as "a prevaricator's fiction is a vacuous thing that provides no handle" for a wh-item as follows: "What kind of thing is a prevaricator's fiction?"

[2]Note: An alternate version of a sentence, clause, or phrase which "means" what another sentence, clause, or phrase "means" is not necessarily a paraphrase according to the rules here presented. Saying a thing in another way is not always equivalent to paraphrasing by these rules.

Such a situation occurs on occasion when a reviewer is dissatisfied with an item stem (or stem plus response) and rewrites the item to make it sound better or to avoid heaviness, awkwardness, wordiness, etc.--but without first writing a new paraphrase or without taking the original paraphrase into consideration. The rewritten item, considered out of context, will often sound or look better, but it will often no longer be an item based on an acceptable paraphrase.

A similar problem arises when an item is rewritten but is no longer a WH-detail item.

[3]Modifiers include adjectives and adverbs, <u>not</u> articles or determiners.

[4]Superordinate terms are not necessarily acceptable synonyms (e.g., <u>dog</u> is not necessarily an acceptable synonym for Siberian wolf-hound).

[5] See Rules for Constructing WH-Detail Items, on file with BSCR.

[6]Average number of WH-detail items written for each passage was more than ten, of which six were selected. Criteria for selection were quality (e.g., absence of awkwardness and turgidity) and freedom from mutual cueing, defined as a stem giving away a response to another stem. In the following, for example, stem A cues the answer to stem B: "A. When did the fuel drums burst into flame?" "B. What burst into flame?"

[7]Fewer than ten per cent of the passages from the original sample had to be replaced.